Automatic ontology-based knowledge extraction from web documents

Journal Item

Automatic Ontology-based Knowledge Extraction and Tailored Biography Generation from the Web

Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal
Wendy Hall, Paul H. Lewis, Nigel R. Shadbolt
Intelligence, Agents, Multimedia Group
University of Southampton

Abstract
This paper presents recent developments in the Artequakt project which seeks to automatically extract knowledge about artists from the Web, populate a knowledge base, and use it to generate personalized narrative biographies. An overview of the system architecture is presented and the three key components of that architecture are explained in detail, namely knowledge extraction, information management and biography construction. An example experiment is detailed and further challenges are outlined.

Keywords: knowledge extraction, ontology population, narrative generation.

Introduction

Accessing and extracting knowledge from Web documents is crucial for the realisation of the Semantic Web and the provision of advanced knowledge services. Even though Web page annotations could facilitate acquiring such knowledge, annotations are rare and in the near future will probably not be rich or detailed enough to cover all the knowledge contained in these documents. Manual annotation is impractical and unscalable, while automatic annotation tools are still in their infancy. Hence specialised knowledge services may require tools able to search and extract specific knowledge directly from unstructured text on the Web, guided by an ontology that details what type of knowledge to harvest.

The use of ontologies to support knowledge extraction has been previously demonstrated (Handschuh et al 2002, Vargas Vera et al 2001). Nevertheless the full potential of this approach is not yet explored. To this end, the Artequakt project aims to dynamically link a knowledge extraction tool with an ontology to achieve continuous knowledge support and guidance to the extraction mechanism. The ontology can provide a domain knowledge classification in the form of concepts and relations. The extraction tool searches online documents and extracts the knowledge that matches the given classification structure, and provides it in a machine-readable format to be automatically maintained in a Knowledge Base (KB). Knowledge extraction could further be enhanced with a lexicon-based term expansion mechanism to enrich the extraction process with extended ontology terminology.

There exist many information extraction (IE) systems that enable the recognition of entities within documents (e.g. 'Rembrandt' is a 'Person', '15 July 1606' is a 'Date'). However, such information is of little use without acquiring the relation between these entities (e.g. 'Rembrandt' was born on '15 July 1606'). Extracting such relations automatically is difficult, but crucial to complete the acquisition of knowledge fragments and ontology population. In the Artequakt project, we attempt to identify entity relationships following ontology relation declarations and lexical information.

Storing information in a structured KB provides the needed infrastructure for a variety of knowledge services. One interesting service is to reconstruct the original source material in new ways, producing a dynamic presentation tailored to the users needs. Previous work in this area has highlighted the difficulties of maintaining a rhetorical structure across a dynamically assembled sequence (Rutledge et al 2000). Where dynamic narrative is present it has been based around robust story-schema such as the format of a news programme (a sequence of atomic bulletins) (Lee et al 1999). It is our belief that by building a story-schema layer on top of an ontology we can create dynamic stories within specific domain. By populating the ontology through automatic knowledge acquisition software we allow those stories to be constructed from the vast wealth of information that exists on the Web.

The Artequakt Project

The Artequakt project aims to implement a system that searches the Web and extracts knowledge about artists, based on an ontology describing that domain, and stores this knowledge in a KB to be used for automatically producing tailored biographies of artists.

The expertise and experience of three separate projects are drawn together under the umbrella of the Artequakt project. These are:

- *The Artiste project* - A European project to develop a distributed database of art images. This has recently been succeeded by *Scultpeur*, which will extend the database to 3-D objects and integrate with the Semantic Web.
- *The Equator IRC* - An EPSRC funded Interdisciplinary Research Collaboration that includes the use of narrative techniques in information structuring and presentation.
- *The AKT IRC* - An EPSRC funded Interdisciplinary Research Collaboration looking at all aspects of the knowledge lifecycle.

The first stage of this project consisted of developing an ontology for the domain of artists and paintings. A selection of IE tools and techniques were developed and applied that attempt to automatically populate the ontology with information extracts from online documents based on the given ontology's representations and WordNet lexicons. The information is stored in a KB and analysed for duplications. In the second stage, narrative construction tools were developed to query the KB through an ontology-server to search and retrieve relevant facts or textual paragraphs and generate a specific biography.

The automatic generation of tailored biographies is concerned with two areas of focus. Firstly, providing biographies for artists where there is sparse information available, distributed across the Web. This may mean constructing text from basic factual information gleaned, or combining text from a number of sources with differing interests in the artist. Secondly, the project aims to provide biographies that are tailored to the particular interests and requirements of a given reader.

To provide a focus for the project and a corpus of data for a demonstrator, the subject domain of impressionist artists and their paintings was chosen. However, the techniques being developed could be applied to other domains.

Architecture Overview

Figure 1 illustrates Artequakt's architecture. Three key areas can be identified. The first concerns the knowledge extraction tools used to extract factual information items together with sentences and paragraphs from Web documents that might be manually selected or obtained automatically using appropriate search engine technology. The fragments of information are passed to the ontology server along with metadata derived from the vocabulary of the ontology. The second key area is the information management and storage. The information is stored by the ontology server and consolidated into a KB which can be queried via an inference engine. The final key area is the narrative generation. The Artequakt server takes requests from a reader via a simple Web interface. The reader request will include an artist of whom a biography to be generated in a particular style (chronology, summary, etc.) and also any user interest, for example the narrative might be generated specifically about the artist's style or paintings. The server uses story templates to render a narrative from the information stored in the KB.

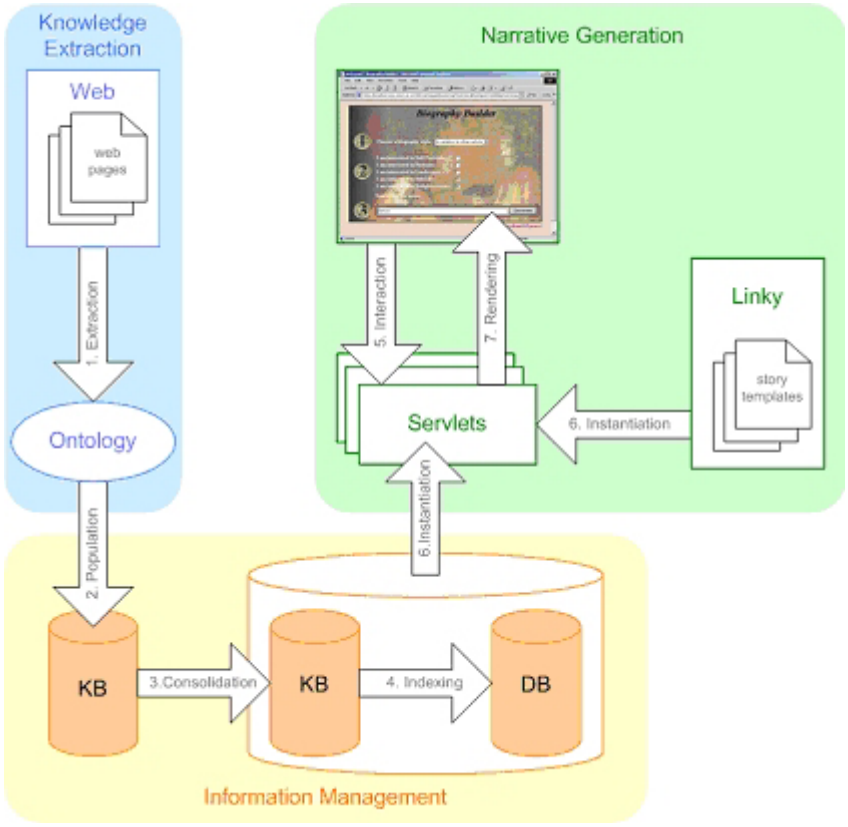

Figure 1. The Artequakt architecture

The Artequakt Ontology

An ontology is a conceptualisation of a domain into a machine readable format. For Artequakt the requirement was to build an ontology to represent the domain of artists and artefacts. This ontology was implemented in Protégé[1], which is a graphical ontology editing tool and is also used to store the knowledge base. The main part of this ontology was constructed from selected sections in the CIDOC Conceptual Reference Model (CRM[2]) ontology. CRM was developed by ICOM/CIDOC Documentation Standards Group to represent an ontology for cultural heritage information. It was built to facilitate the transformation of existing disparate museum and cultural heritage information sources into one coherent source.

The CRM ontology is designed to represent artefacts, their production, ownership, location, etc. This ontology was modified for Artequakt and enriched with additional classes and relationships to represent a variety of information related to artists, their personal information, family relations, relations with other artists, details of their work, etc. The Artequakt ontology also allows the storage of textual paragraphs and sentences along with their source URLs so that at a later point they can be reorganised using the ontology as a guide.

Knowledge Extraction

Much of information on the Web is in the form of natural language documents. A promising approach to accessing the knowledge in such documents is centred on IE that reduces the documents to tabular structures from which the fragments of documents can be retrieved as answers to queries. However, the time and effort needed for manually annotating a large number of texts and the prerequisite of templates that stipulates which types of information

[1] http://protege.stanford.edu/
[2] http://cidoc.ics.forth.gr/index.html

are extractable are major challenges of exploiting such extraction techniques for practical purposes (Yangarber et al 2001).

Many IE systems rely on predefined templates and pattern-based extraction rules or machine learning techniques in order to identify certain entities within text documents. Documents on the Web use limitless vocabularies, structures and composition styles for defining approximately the same content, making it hard for any IE technique to cover all variations of writing patterns. For example, although content similarity between two biographic documents might be expected, expressions used for both sources may vary significantly. More importantly, traditional IE systems lack the domain knowledge required to pick out relationships between the extracted entities.

These observations lead us to the use of an ontology coupled with a general-purpose lexical database (WordNet[3]) and an entity-recogniser (GATE[4]) as guidance tools for identifying knowledge fragments consisting of not just entities, but also the *relations* between them. Automatic term expansion is used to increase the scope of text analysis to cover syntactic patterns that imprecisely match our definitions.

Extraction Procedure

Figure 2 is the prototype user interface which allows the user to search for an artist and select the profile of interest. When the user enters an artist name, e.g. *Rembrandt*, a quick search in the KB is made to check if biographical knowledge for this artist already exists.

Figure 2. Artequakt's user interface.

If the given artist is new to the KB, a script will be deployed to search the Web using the artist's name as input. It is necessary to filter out irrelevant documents from the search engine results. We have explored the idea of 'searching by example' that uses example documents from trusted sites as a basis for measuring similarity between the query and the search results. Exemplars are obtained from the Web Museum[5] site that provides short descriptions of artists. A vector similarity measure is applied to compare the Web Museum page that describes the given artist with the search engine results, selecting only those with a similarity above a

[3] Available at http://www.cogsci.princeton.edu/~wn
[4] Available at http://gate.ac.uk
[5] http://www.ibiblio.org/wm/paint

certain threshold. The left side of Table 1 shows some of the accepted and rejected Web sites about *Rembrandt*. The accepted sites provide a variety of information about Rembrandt's life and paintings, and were filtered out of more than 60 URLs returned by *Yahoo* and *AltaVista* search engines. The rejected sites listed in the table are irrelevant as they represent restaurants and hotels named Rembrandt.

The disadvantage of this filtering method is evident in that it is not always possible to locate a good example document, especially when the query is for a relatively new, or unknown artist. One possible improvement is to expand queries with related terms to improve the search results, but this will unlikely remove the need for a filtering process. We are currently experimenting with measuring document relevancy with respect to term vectors constructed directly from the ontology terminology.

Accepted Site	***Rejected Sites***
http://www.ibiblio.org/wm/paint/auth/rembrandt/	http://www.rembrandt-s.com/
http://www.mcs.csuhayward.edu/~malek/Rembran.html	http://www.rembrandts.com/
http://www.artchive.com/artchive/R/rembrandt.htmls	http://www.hotelrembrandt.co.uk/

Table 1. Filtered set of Web documents about Rembrandt.

After the selection of documents is made, each selected Web document is then divided into paragraphs, which are in turn broken down into sentences. Each paragraph is analysed syntactically and semantically to identify any relevant knowledge to extract. The Apple Pie Parser[6] is used for grouping grammatically related phrases as the result of syntactical analysis. Semantic examination then locates the main components of a given sentence (i.e. 'subject', 'verb', 'object'), and identifies named entities (e.g. 'Rembrandt' is a 'Person', 'Leiden' is a 'Place') using GATE and WordNet. GATE is also used to resolve anaphoric references (personal pronouns).

The following is an example paragraph extracted from the first selected Web document:

> "Rembrandt Harmenszoon van Rijn was born on July 15, 1606, in Leiden, the Netherlands. His father was a miller who wanted the boy to follow a learned profession, but Rembrandt left the University of Leiden to study painting. His early work was devoted to showing the lines, light and shade, and color of the people he saw about him."

The challenge now is to extract binary relationships between any identified pair of entities. Knowledge about the domain specific semantics is now required, which can be inferred from the ontology and used to decide which relations are required and expected between the entities in hand. At this stage, Artequakt submits a query to the ontology server to obtain such knowledge. In addition, three lexical chains (synonyms, hypernyms, and hyponyms) from WordNet are used in order to reduce the problem of linguistic variation between relations defined in the ontology and the extracted text. For example, the concept of 'depict' can be matched with 'portray' (synonym) and 'represent' (hypernym). Since a relation may have multiple entries in WordNet (polysemous words), the mapping between a relation and an entry in WordNet takes into account syntactic and semantic clues present in a sentence. For example, the relation of *date_of_birth* is mapped into the concept of 'birth' which, according to WordNet, has four noun senses and one verb sense. The 1st noun sense is selected since one of its hypernyms is 'time period' which had 'Date' as a hyponym. The extracted synonyms for the verb sense are 'give birth' and 'bear'. By providing the IE process with direct access to the concepts and relations in the ontology, our approach bypasses the need for predefining external templates.

[6] Available at http://www.cs.nyu.edu/cs/projects/proteus/app/

Figure 3 shows the extraction results based on the sentence "*Rembrandt Harmenszoon van Rijn was born on July 15, 1606, in Leiden, the Netherlands*" found in the example paragraph above. Annotations provided by GATE and WordNet highlight that "Rembrandt Harmenszoon van Rijn" is a person's name, "July 15, 1606" is a date, and "Leiden" and "Netherlands" are locations. Relation extraction is determined by the categorisation result of the verb 'bear' which matches with two potential relations; 'date_of_birth' and 'place_of_birth'. Since both relations are associated with "July 15, 1606" and "Leiden" and "Netherlands" respectively, this sentence generates two knowledge triples about *Rembrandt*:

```
Rembrandt - date_of_birth - July 15, 1606
Rembrandt - place_of_birth - Leiden, Netherlands
```

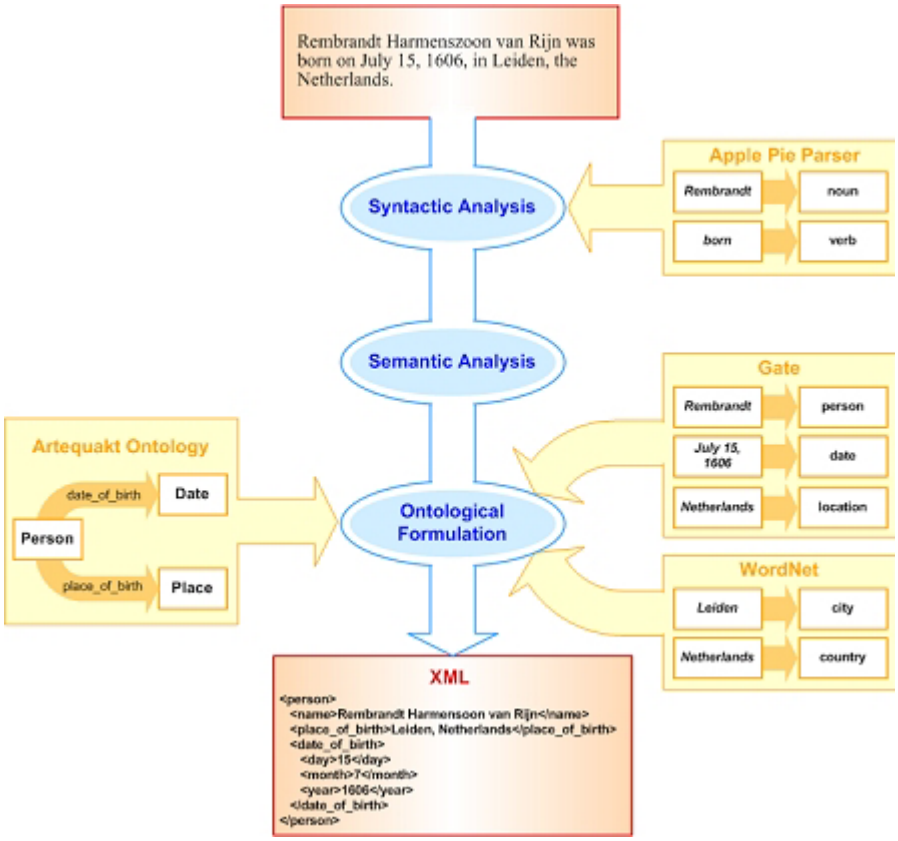

Figure 3. An example of knowledge extraction

The output from each extraction process is an XML representation of the facts, paragraphs, sentences and keywords identified in the selected documents. The extraction process terminates by sending the new XML files to the ontology server to be inserted into the KB.

Currently the knowledge acquisition process is only launched when no information is available in the KB about the query artist. This will soon be extended to allow the user to submit a specific URL for analysis, or to request a new search if the biographies presented do not contain the required information. Furthermore, the extraction process could be reinitiated periodically, searching for new Web sites to update or add to the information about artists that already exist in the KB.

Automatic Ontology Population

Populating ontologies with a high quantity and quality of instantiations is one of the main steps towards providing valuable and consistent ontology-based knowledge services. Manual ontology population is very labour intensive and time consuming. Some semi-automatic approaches have investigated creating document annotations and storing the results as assertions in an ontology. For example in Vargas Vera et al (2001), relationships were added automatically between instances only if these instances already exist in the KB, otherwise user intervention is needed. Handschuh et al (2002) describe a framework for user-driven ontology-based annotations, enforced with the IE learning tool; Amilcare (Ciravegna et al 2002). However, the framework lacks the capability of identifying relationships reliably.

In Artequakt we investigate the possibility of moving towards a fully automatic approach of feeding the ontology with knowledge extracted from the Web. Information is extracted in Artequakt with respect to a given ontology (e.g. the artist ontology described earlier), and provided as XML files, one per document, using tags mapped directly from names of classes and relationships in that ontology. Figure 4 (a) shows an example of the XML representation of the extracted knowledge, and (b) how it is asserted in the ontology.

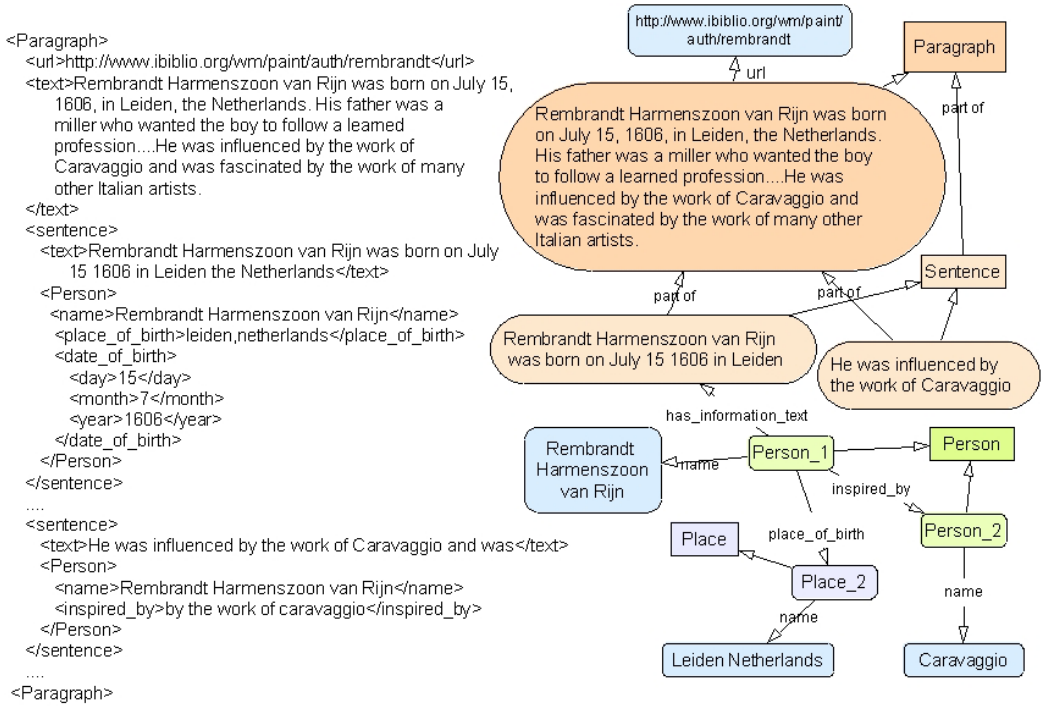

Figure 4. a) XML file of extracted information, b) The corresponding instances and relationships in the ontology.

The ontology server we are experimenting with is based on Java sockets and connected to the Artequakt KB through the Protégé API. A limited inference engine is built on this server to allow querying the KB to retrieve specific information, for example to get all paragraphs that mention the date of birth of a specific artist, get the artist of a painting, get all available facts about an artists, etc. Artequakt's ontology server sends some of the extracted knowledge to a relational database to provide fast access to frequently used information via SQL queries when generating biographies.

Narrative Generation

While machines could benefit from using structured ontologies to exchange information, human beings need a more intuitive interface. One of the most natural ways to do this is by story telling. There is a wealth of critical and philosophical thought concerning narrative that can be drawn on to assist in constructing a story (in this case a biography) from the raw information gathered.

Narratologists frequently divide the subject of study into two, often termed *story* and *discourse*. Story represents the basic description of the fundamental events, while *discourse* refers to the techniques used to vary the presentation of the basic story.

In Artequakt, the KB can be thought of as containing our underlying story. Abstract pieces of information are stored in the knowledge base without any explicit ordering, just relationships. Strictly speaking, because the fragments may be paragraphs of text harvested from Web pages, they may already contain elements of discourse (focalisation, tense information etc). To produce the eventual discourse (in our case pages of html) we need to first arrange sub-elements of the story into a sensible sequence and then render them into an actual text.

Biography Templates

The structures we use to arrange the story are human authored biography templates that contain queries into the data and KBs. The templates are authored in the Fundamental Open Hypermedia Model (FOHM) and stored as XML in the Auld Linky contextual structure server (Michaelides et al 2001).

Any given biography is constructed from several sub-structures. The basic structure used is the *Sequence*. This represents a list of queries that have to be instantiated and inserted into the biography in order. These queries are authored using the vocabulary of terms defined within the ontology. Other structures allow more complex effects. A *Concept* structure contains several queries, any of which may be used at this point in the biography. A *Level of Detail* (LOD) structure is similar to a concept, but there is an ordering between the queries that corresponds to preference (i.e. preferably the highest numbered query should be used, if that's not possible the next highest, and so on). These structures may be nested (e.g. a sequence of concepts).

Some queries retrieve paragraphs directly while others query the KB for specific facts and construct sentences dynamically from the results. This can be useful for facts that have been inferred (and therefore there is no corresponding paragraph), or when there is no paragraph that fits the literary form of the rest of the biography (e.g. the biography is in third person, but all the available paragraphs are in first person).

Figure 5 shows an example template structure. In this case there is a sequence of four story fragments. These can be either database queries (that are resolved into an original paragraph) or KB queries (sentences that must be constructed). The fourth entry in the sequence is an XOR choice (implemented by a LOD structure): if no paragraph can be found then the sentence will be constructed from the knowledge stored in the KB.

The templates also contain contextual information on which parts of the biography structure are appropriate in different contexts. When a user queries Auld Linky for the template, they specify their context and the inappropriate parts of the template are pruned away.

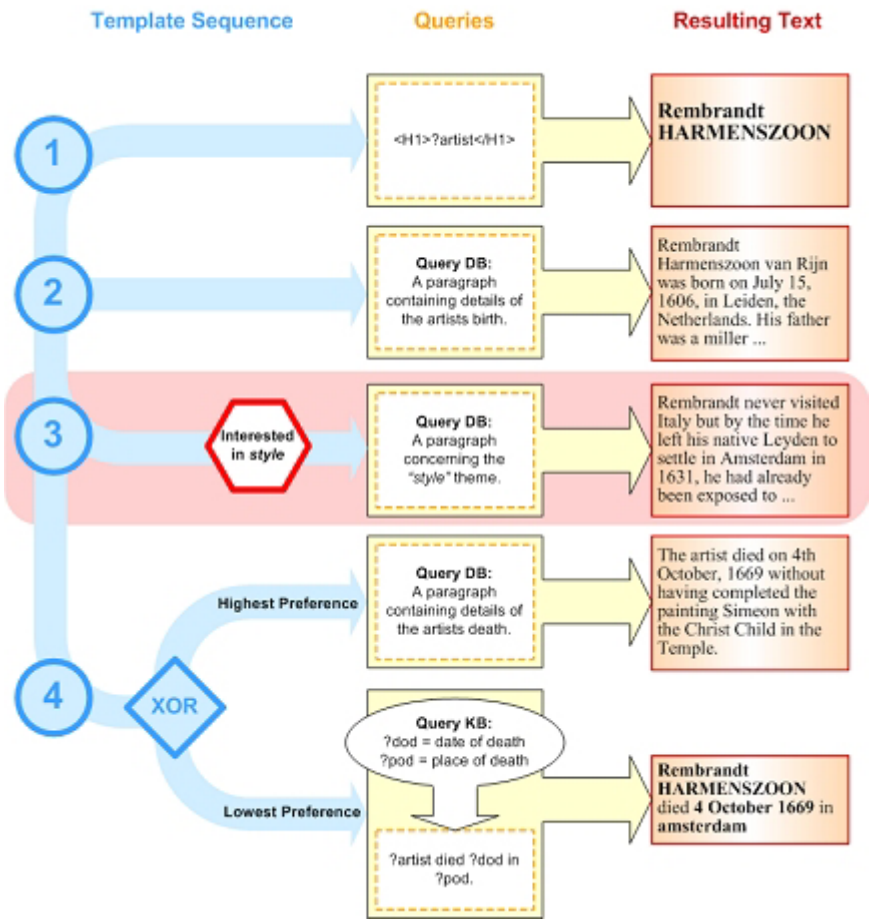

Figure 5. Templates contain queries that are resolved into the final text

For example in Figure 5 the third item in the sequence, concerning the style of the artist, is marked up as being only suitable for those interested in this topic, otherwise the entire third branch would be ignored. In this way the biography structures are tailored to the user's interests.

Once it has been retrieved from Auld Linky, the template has to be instantiated by making each query in turn and then rendering the results into an html page for display. Figure 6 shows a biography on Rembrandt, which is the final result of our example template (where context specifies an interest in style). The heading has been generated by taking the instance name of the artist from the KB. Templates may include their own text; the sub-heading *Summary Biography* is an example of this. The first paragraph concerns the birth of Rembrandt and was extracted from the Web Museum site. The second section is comprised of a number of pieces of text from different sites that all describe Rembrandt's painting style.

The final sentence provides details of Rembrandt's death. No suitable paragraph existed in the DB, however the basic facts were present in the KB and these were used to construct a simple sentence.

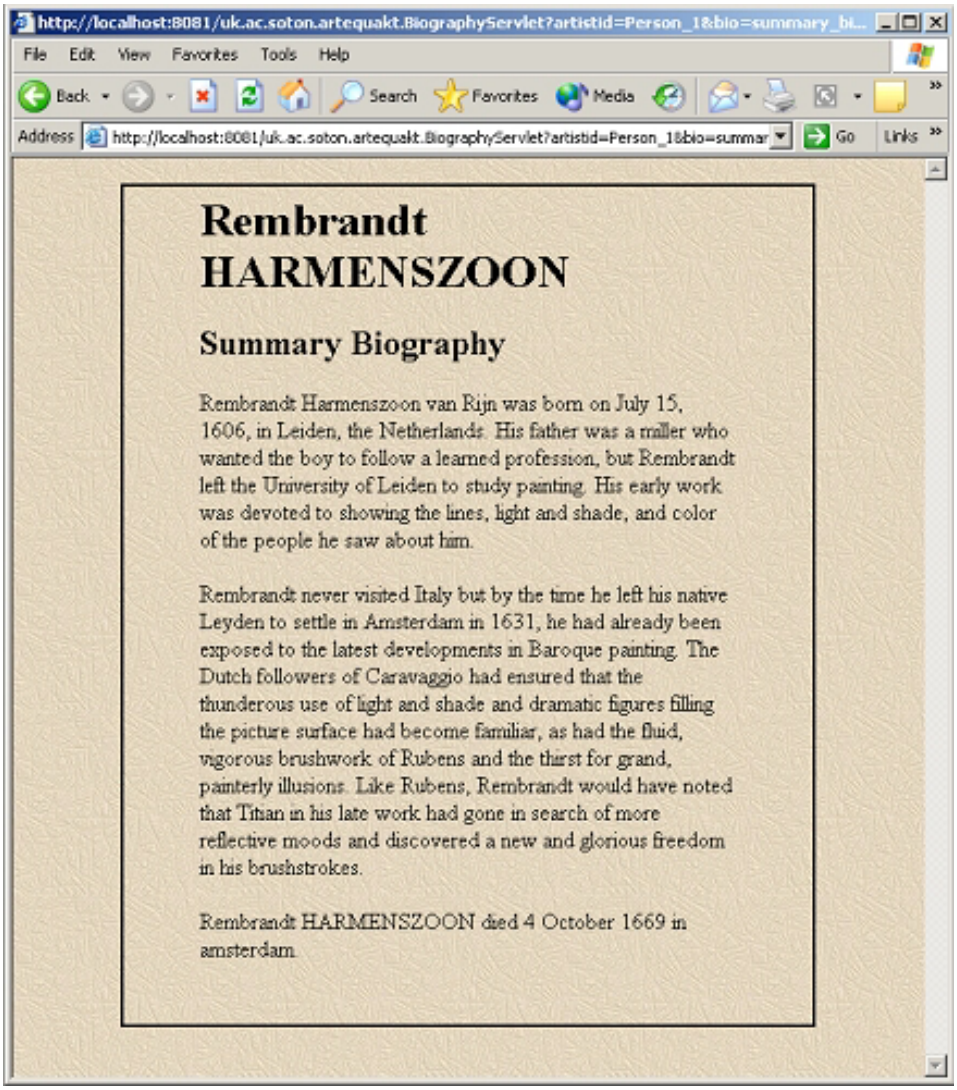

Figure 6. The final rendered biography

Summary

The system we produced integrates a variety of tools in order to automate an ontology-based knowledge acquisition process and maintain a KB which is used to generate customised biographies.

Work on Artequakt included building a generic knowledge extraction tool that is dynamically integrated to an ontology. The ontology supplies the needed knowledge about the domain and type of information to extract, and how to represent this information in a proper metadata format to insert automatically into a specific KB. We used GATE and WordNet for entity identification. WordNet was also used to expand the terms in the ontology and Web documents. Many entity relationships were successfully extracted and used later on to construct biographies.

The running example described in this paper gives a step-by-step demonstration of Artequakt's functionality, illustrating how and what each tool contributes to the working system.

Further Challenges

The Artequakt project is dealing with many challenging issues related to ontology population and maintenance, knowledge extraction, and the generation of personalised artist biographies.

User-driven ontology population tools need to deal with the issues of duplicate information across documents and redundant annotations (Staab et al 2001). The same problems can be encountered with automatic ontology population approaches. Even though such approaches help to speedily acquire large quantities of knowledge, precision and recall may decrease. Automatically populating an ontology from diverse and distributed Web resources poses significant challenges. One persistent and prevalent problem is that of the consolidation of duplicate information that arises when extracting similar or overlapping information from different sources. Tackling this problem is important to maintain the referential integrity and quality of results of any ontology-based knowledge service (Alani et al 2002). A description of the Artequakt's approach to tackle these problems can be found in Kim et al (2002).

The Artequakt extraction tool retrieves metadata triplets (subject – relation – object) from the ontology automatically to overcome the limitations of predefined fixed templates. This flexibility is increased when the ontology class and relationship names are minimal, thus avoiding, for example, compound or obscure names that may confuse concept identification and term expansion. The approach presented in this paper has shown good potential for covering semantic variations. However, the coverage may be decreased when the term expansion is limited. In addition, the extraction tool will not be able to distinguish between ontology relations with synonymous names, connecting the *same* classes. Such scenarios can be avoided if appropriate synonyms are specified for the classes and relations in the ontology at the design stage in order to remove possible ambiguities.

One other challenge with automatic extraction is specificity. For example, it is not very difficult to identify an entity as a Person, while it is much harder to tell whether this person is a Painter, or a Sculptor. That level of knowledge could be inferred when more facts are extracted about that person, e.g. information about paintings or sculptures produced.

Generating a narrative also provides significant challenges. Extracted sentences may often contain coreferences that need to be resolved in the context of the whole narrative. For example, a sentence may use the pronoun *he* that only makes sense if the artist was the previous subject. Similarly replacing occurrences of the artists name with appropriate pronouns may make the text more readable.

When rendering the narrative, problems of duplication may occur when selected sentences contain more information than requested by the template. For example, a query for a sentence containing, for example, a date of birth may include information on the artist's parents. If the system keeps track of this additional information, later duplication could be avoided. On the other hand, there may exist many paragraphs concerning the artist's family life. Determining how many of such paragraphs to include in the biography may depend on factors including the desired overall length of the biography, the amount of interest in that topic, etc. By tackling these challenges in narrative generation, we can improve the quality of the biographies produced.

The Artequakt approach should, in theory, be applicable to other domains with little technical change. For example, the current artist ontology may be replaced with an actor ontology, where the extraction is expected to focus on information about actors, as guided by the new ontology. In addition, with respect to entity recognition performance, domain-specific entities (e.g. painting styles, films) need specialised extraction rules that have to be modified when the domain changes. Future work will investigate and develop further the generic architecture of the Artequakt system.

Acknowledgement

The work presented here is part of a larger project and we would particularly like to note the contributions of Hugh Glaser, Srinandan Dasmahapatra and David De Roure. This research is funded in part by EU Framework 5 IST project "Scultpeur" IST-2001-35372, EPSRC IRC project "Equator" GR/N15986/01 and EPSRC IRC project "AKT" GR/N15764/01.

References

H. Alani, S. Dasmahapatra, N. Gibbins, H. Glaser, S. Harris, Y. Kalfoglou, K. O'Hara and N. Shadbolt, "Managing Reference: Ensuring Referential Integrity of Ontologies for the Semantic Web," *Proc. 13th Int'l Conf. Knowledge Eng. and Knowledge Management* (EKAW'02), Lecture Notes in Artificial Intelligence, Siguenza, Spain, 2002.

F. Ciravegna, A. Dingli, Y. Wilks, and D. Petrelli, "Timely and Non-Intrusive Active Document Annotation via Adaptive Information Extraction," *Workshop on Semantic Authoring, Annotation & Knowledge Markup* (SAAKM'02), *15th European Conference on Artificial Intelligence* (ECAI'02), Lyon, France, 2002, pp. 7-13.

S. Handschuh, S. Staab, and F. Ciravegna, "S-CREAM - Semi-Automatic Creation of Metadata," *Semantic Authoring, Annotation and Markup Workshop, 15th European Conference on Artificial Intelligence*, (ECAI'02), Lyon, France, 2002, pp. 27-33.

S. Kim, H. Alani, W. Hall, P. H. Lewis, D. E. Millard, N. Shadbolt, and M. J. Weal, "Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web," *Workshop on Semantic Authoring, Annotation & Knowledge Markup* (SAAKM'02), *the 15th European Conference on Artificial Intelligence*, (ECAI'02), Lyon, France, 2002, pp. 1-6.

K. Lee, D. Luparello, and J. Roudaire, "Automatic Construction of Personalised TV News Programs," *Proc. 7th ACM Conf. on Multimedia*, Orlando, Florida, 1999, pp. 323-332.

D.T. Michaelides, D.E. Millard, M.J. Weal, and D. DeRoure, "Auld Leaky: A Contextual Open Hypermedia Link Server", *Hypermedia: Openness, Structural Awareness, and Adaptivity, Proc. OHS-7, SC-3, and AH-3*, Lecture Notes in Computer Science (LNCS 2266), Springer Verlag, Heidelberg, 2001, pp.59-70.

L. Rutledge, B. Bailey, J.V. Ossenbruggen, L. Hardman, and J. Geurts, "Generating Presentation Constraints from Rhetorical Structure," *Proc. 11th ACM Conf. on Hypertext and Hypermedia*, San Antonio, Texas, USA, 2000, pp. 19-28.

S. Staab, A. Maedche, and S. Handschuh, "An Annotation Framework for the Semantic Web," *Proc. of the 1st Int'l Workshop on MultiMedia Annotation*, Tokoyo, Japan, 2001.

M. Vargas-Vera, E. Motta, J. Domingue, S. Buckingham Shum, and M. Lanzoni, "Knowledge Extraction by using an Ontology-based Annotation Tool," *Proc. First Int'l. Conf. on Knowledge Capture*, (K-CAP'01), *Workshop on Knowledge Markup & Semantic Annotation*, Victoria, B.C., Canada, 2001.

R. Yangarber, R. Grishman, "Machine Learning of Extraction Patterns from Unannotated Corpora: Position Statement," *Proc. Workshop on Machine Learning for Information Extraction*, Berlin, 2001, pp. 76-83.